KLASIFIKASI FITUR DALAM DOKUMEN *REVIEW* PRODUK DENGAN METODE *LOCAL POINTWISE MUTUAL INFORMATION*

Yufis Azhar
Program Studi Teknik Informatika
Fakultas Teknik, Universitas Muhammadiyah Malang
Email : yufis.az@gmail.com

ABSTRAK

Ekstraksi fitur produk dalam suatu dokumen review merupakan permasalahan yang telah menarik perhatian banyak peneliti untuk memecahkannya. Permasalahan utama pada topik ini adalah bagaimana mengekstrak fitur yang relevan. Salah satu cara yang umumnya dilakukan adalah dengan mengkategorikan fitur-fitur yang telah terekstrak ke dalam kelas-kelas tertentu. Akan tetapi metode klasifikasi yang digunakan biasanya memiliki akurasi cukup rendah, hal ini dikarenakan sifat dokumen opini yang sangat bergantung pada domain yang sedang dibicarakan, Sehingga metode klasifikasi yang digunakanpun harus mampu beradaptasi dengan sifat tersebut. Dalam penelitian ini diusulkan suatu metode Local Pointwise Mutual Information (LPMI) yang merupakan modifikasi dari teknik PMI yang selama ini digunakan. Letak perbedaan utamanya adalah pada area pencarian PMI yang bersifat local (hanya di dataset) sehingga tidak keluar dari domain yang sedang dibicarakan oleh pemberi opini. Hasil pengujian menunjukkan bahwa teknik ini memiliki nilai precision dan recall yang baik dengan rata-rata di atas 80%.

Kata Kunci: *feature-based opinion mining, ekstraksi fitur, klasifikasi fitur, pointwise mutual information.*

Abstract

Extraction of product features in a review document is a problem that has attracted the attention of many researchers to solve it. The main issues on this topic is how to extract the relevant features from the review. The popular method is categorize the features that have been extracted into certain classes. But the classification method used usually have low accuracy. This is happen because the review documents are domain dependent. This research proposed a Local Pointwise Mutual Information (LPMI) to solve the problem. This method is a modification from the traditional PMI method. The search area of LPMI is only inside the review documents from the same domain. That is the main difference from the traditional PMI method. The test result show that this technique has a average recall and precision value above 80%.

Keyword: *feature-based opinion mining, feature extraction, feature classification, pointwise mutual information*

1. PENDAHULUAN

Ketersediaan dokumen komentar atau opini dengan jumlah besar di internet saat ini menarik perhatian beberapa orang peneliti untuk menggali informasi yang terkandung di dalamnya. Bidang ini selanjutnya dikenal dengan istilah *opinion mining*. Pengekstrakan target opini (*feature based opinion mining*) adalah salah satu masalah yang seringkali diangkat oleh para peneliti. Pada suatu teks opini, target ini biasanya berupa frase/kata benda [1]. Sebagai contoh dalam kalimat "*Canon S100 has a great lens*", target opininya adalah "*lens*". Dalam teks-teks opini tentang suatu produk yang tersebar di internet, target opini ini seringkali ditujukan bukan hanya untuk produk, tetapi juga untuk fitur (komponen) dari produk tersebut. Oleh karena itu, dalam dokumen *review* suatu produk, yang menjadi target opini biasanya adalah fitur dari produk itu sendiri.

Permasalahan bagaimana mendapatkan fitur produk dari suatu teks opini cukup kompleks. Harus ada suatu *identifier* yang dapat mengenali kata benda yang merupakan fitur produk. Hu mengatakan bahwa kata sifat yang memiliki hubungan dengan kata benda dapat menjadi identifier yang kuat untuk mengenali target opini dalam suatu kalimat [1]. Qiu mengusulkan algoritma *double propagation* yang merupakan metode *semi unsupervised* untuk mengekstrak fitur produk [2]. Metode ini tidak membutuhkan kamus kata sifat secara lengkap, karena metode ini dapat melengkapi kamus katanya secara otomatis. Caranya adalah dengan menemukan fitur

menggunakan kamus kata sifat, kemudian memanfaatkan fitur yang telah terekstrak tadi untuk menemukan kata sifat lain yang terdapat dalam teks opini. Kata sifat baru tersebut secara otomatis akan ditambahkan dalam kamus kata. Proses ini berlangsung terus-menerus hingga tidak ada fitur dan kata sifat baru yang ditemukan. Akan tetapi, metode *double propagation* yang diusulkan oleh Qiu tersebut memiliki beberapa kelemahan, diantaranya adalah ketidakmampuan metode tersebut untuk mengenali fitur produk yang dirujuk oleh suatu kata ganti. Permasalahan ini dapat diselesaikan dengan penambahan rule yang dapat mengenali kata ganti (*pronoun*) yang memiliki hubungan ketergantungan dengan suatu kata sifat serta merujuk pada suatu kata benda (*noun*) [3].

Masalah lain yang belum dapat diselesaikan oleh metode *double propagation* adalah terlalu banyaknya fitur produk yang terekstrak. Beberapa fitur produk sejatinya merujuk pada objek yang sama sehingga sebenarnya bisa digabung. Sebagai contoh kata *picture* dan *photo* sebenarnya merujuk pada objek yang sama. Dalam kasus yang lain, masalah tersebut mungkin bisa diselesaikan dengan menggunakan sinonim. Akan teteapi untuk kasus pengenalan fitur produk, sinonim tidak dapat digunakan karena sifat kalimat opini yang sangat bergantung pada *domain*. Sebagai contoh dalam dokumen review film, kata *picture* dan *movie* merujuk pada objek yang sama, padahal kedua kata tersebut bukan sinonim. Rozi berusaha untuk memecahkan masalah ini dengan melakukan *clustering* terhadap fitur produk yang berhasil diekstrak, kemudian melakukan pelabelan secara otomatis terhadap *cluster* yang terbentuk [4]. Akan tetapi cara inipun tidak optimal karena banyak label dari *cluster* yang justru tidak relevan dengan prodek yang di *review*.

Berdasarkan permasalahan tersebut, dalam penelitian ini diusulkan suatu metode untuk klasifikasi fitur produk yang dapat mengelompokkan fitur produk berdasarkan kedekatan makna dengan tetap memperhatikan sifat ketergantungan pada *domain* yang dimiliki suatu kalimat opini.

2. DASAR TEORI

2.1. METODE *DOUBLE PROPAGATION*

Metode double propagation adalah metode semi *unsupervised* yang diusulkan oleh Qiu [2]. Pada dasarnya, metode ini akan mengekstrak kata opini (atau target opini) berulang kali menggunakan kata opini (atau target opini) yang telah diketahui atau telah terekstrak sebelumnya melalui identifikasi relasi *syntactic* nya. Metode ini disebut *semi unsupervised* karena masih membutuhkan bantuan dari *opinion lexicon* (kamus kata opini). Akan tetapi kamus ini tidak harus lengkap, karena dalam prosesnya kamus kata ini akan dilengkapi secara otomatis.

Terdapat 4 masalah utama dalam pengekstrakan kata opini dan target opini (dalam hal ini, target opini adalah fitur produk). yang harus ditangani oleh *double propagation* (DP), yaitu : bagaimana mengekstrak target opini menggunakan kata opini; bagaimana mengekstrak target opini menggunakan target opini; bagaimana mengekstrak kata opini menggunakan target opini; dan bagaimana mengekstrak kata opini menggunakan kata opini. Seperti dapat dilihat dalam keempat permasalahan tersebut, Qiu memfokuskan pengamatan pada kemunculan kata benda (*noun*), yang dianggap sebagai fitur produk, dan kata sifat (*adjective*), yang dianggap sebagai kata opini, pada suatu kalimat. Berdasar pada hal tersebut, Qiu menyusun aturan-aturan yang digunakan untuk mengekstrak kata opini dan target opini dalam suatu kalimat.

Metode ini dapat dikatakan sebagai metode yang mendekati sempurna untuk mengekstrak fitur/target opini. Karena metode DP tidak membutuhkan kamus kata yang lengkap dan juga mampu menghilangkan masalah *dependency* dalam suatu dokumen opini. Akan tetapi, metode DP memiliki suatu kelemahan yaitu masih banyaknya fitur yang tidak relevan yang ikut terekstrak. Fitur-fitur yang tidak relevan tersebut ikut terekstrak karena memenuhi aturan yang didefinisikan oeh DP, salah satunya aturan yang mengatakan bahwa jika terdapat suatu kata benda (*noun*) memiliki hubungan ketergantungan dengan suatu kata sifat (*adjective*), maka kata benda tersebut adalah fitur produk/target opini. Maka ketika terdapat kalimat "*This is the best one*", kata "*one*" yang merupakan kata benda, akan terekstrak sebagai fitur produk karena

memiliki relasi ketergantungan dengan kata "*best*". Permasalahan itulah yang ingin dipecahkan dalam penelitian ini.

2.2. *POINTWISE MUTUAL INFORMATION*

Metode *Pointwise Mutual Information* (PMI) adalah suatu metode yang seringkali digunakan untuk melihat kedekatan antara dua buah kata. Metode ini akan melihat rasio kemunculan kedua kata tersebut secara berbarengan dalam satu dokumen dengan kemunculan kata tersebut secara terpisah salam dokumen yang lain. Seperti yang ditunjukkan oleh persamaan (1)

$$D_{ij} = \frac{F_{ij}}{F_i \times F_j} \quad (1)$$

dimana D_{ij} adalah kedekatan kata i dengan kata j, F_{ij} adalah jumlah dokumen yang mengundang kata i dan kata j, F_i adalah jumlah dokumen yang mengandung kata i, sedangkan F_j adalah jumlah dokumen yang mengandung kata j. Semakin besar nilai D_{ij} maka semakin dekat kata tersebut demikian pula sebaliknya.

Metode ini pernah digunakan oleh Turney untuk mengklasifikasikan suatu kata sifat ke dalam kelas positif dan *negative* dengan cara mencari kedekatan kata tersebut dengan kata *good* (untuk kelas positif) dan kata bad (untuk kelas negatif) [5]. Kata-kata sifat tersebut nantinya akan digunakan untuk mengelompokkan suatu kalimat opini dalam kelas yang sama yaitu positif dan negatif.

3. METODOLOGI PENELITIAN

3.1. *Unified Modeling Language* (UML)

Penelitian ini dititikberatkan pada bagaimana mengoptimasi kinerja dari metode *double propagation* dalam proses ekstraksi fitur produk. Khususnya dalam pengklasifikasian fitur produk yang dihasilkan. Oleh karena itu, untuk mempermudah proses pengerjaan, tahap preprocessing data akan dilakukan dengan menggunakan *library* dari *Stanford Parser* dan *Stanford POS Tagger*.

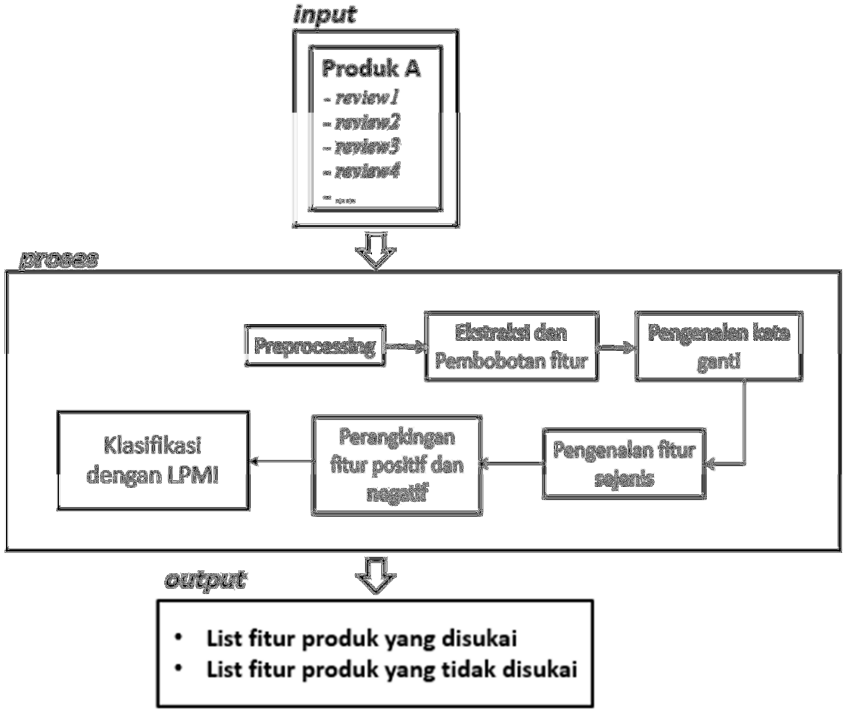
Sedangkan dataset yang digunakan didapat dengan cara crawling dari situs jual beli online berbahasa inggris, yakni amazon.com. Situs ini dipilih karena merupakan salah satu situs jual beli online terbesar dan memiliki customer yang tersebar hampir di seluruh dunia. Dokumen *review* produk didapatkan melalui metode *crawling* dengan memanfaatkan API yang sudah disediakan oleh Amazon. Dengan menggunakan API ini, selain dokumen *review*, juga bisa didapatkan informasi-informasi lain seperti nama produk, rating yang diberikan *customer* untuk tiap produk, id member yang memberikan komentar, skor *helpful feedbacks* dari tiap dokumen *review*, dan lain sebagainya. Produk yang akan diambil dokumen *review* nya adalah 10 produk *smartphone* yang memiliki *range* harga mulai dari $200 hingga $400. *Range* harga ini dipilih karena produk-produk yang berada pada *range* harga tersebut memilik jumlah komentar yang cukup banyak. Jumlah komentar untuk kesepuluh produk tersebut berjumlah 1.610 buah dengan masing-masing komentar memiliki minimal 1 kalimat dan paling banyak 5 kalimat.

Sistem yang dibangun adalah sistem untuk ekstraksi fitur produk yang dikomentari oleh *customer*. Input dari sistem ini adalah dokumen yang berisi daftar komentar untuk setiap produk *smartphone*, sedangkan *output* yang diharapkan adalah daftar fitur yang dimiliki oleh produk tersebut yang dikomentari oleh *customer* (yang telah terklasifikasi). Tujuannya adalah untuk mendapatkan fitur apa saja yang disukai dan tidak disukai oleh *customer*. Alur kerja sistem secara umum dapat dilihat pada Gambar 1.

Dalam penelitian ini, metode DP digunakan untuk ekstraksi fitur produk dalam suatu dokumen opini. Proses ekstraksi ini nantinya akan menghasilkan daftar kata yang dianggap sebagai fitur produk. Daftar fitur produk yang dihasilkan tersebut kemudian diranking berdasarkan frekuensi kemunculannya dalam keseluruhan dokumen. Setelah diranking, fitur-fitur yang memiliku frekuensi kata kurang dari *threshold* yang ditentukan akan dibuang. Proses ini dilakukan untuk mendapatkan fitur penting dari produk tersebut.

Setelah proses tersebut dilakukan, fitur-fitur tadi akan diklasifikasikan menggunakan metode *Local Pointwise Mutual Information* (LPMI). Metode ini merupakan modifikasi dari metode PMI konvensional. Dimana dataset yang dijadikan acuan adalah dataset yang memiliki *domain* yang sama dengan dokumen opini yang sedang dianalisis. Hal ini dilakukan untuk mengatasi masalah ketergantungan *domain* yang seringkali dialami oleh metode klasifikasi yang lain.

Dalam penelitian ini, digunakan 5 kata kunci yang mewakili tiap class dalam fitur produk smartphone, yaitu *Design, Batery, Memory, Camera,* dan *Prize*.



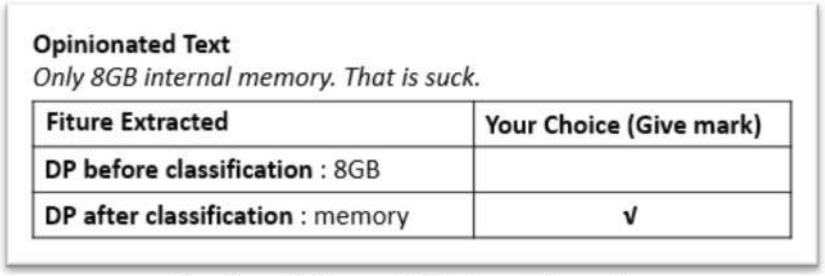
Gambar 1. Alur Kerja Sistem

4. PENGUJIAN DAN PEMBAHASAN

Dalam penelitian ini dilakukan beberapa ujicoba untuk menganalisa kinerja dari sistem. Ada 2 skenario ujicoba yang dilakukan, yaitu :

1. Uji kuisioner dengan melibatkan beberapa orang *user* untuk membandingkan hasil ekstraksi fitur dari metode DP sebelum diklasifikasi, dengan hasil ekstraksi fitur dari metode DP setelah diklasifikasi.
2. Uji *variable precision* dan *recall* untuk melihat akurasi hasil klasifikasi fitur dengan metode *Local Pointwise Mutual Information*.

Untuk pengujian tersebut, sebanyak 30 *user* dilibatkan. *User* ini adalah orang yang tergabung dalam salah satu grup pecinta teknologi *mobile*, sehingga cukup akrab dengan istilah-istilah yang sering digunakan dalam mengomentari suatu fitur dari sebuah *smartphone*. Teknik kuisioner yang dilakukan adalah membagi dataset ke dalam kategori 10 buah produk. Masing-masing *user* diberikan 2 buah kategori produk untuk dianalisa. Sehingga 1 kategori produk akan dianalisa oleh 6 orang *user*. Tugas dari *user* ini adalah membaca opini dari produk tersebut, kemudian *user* membandingkan fitur produk yang berhasil diekstrak oleh metode DP sebelum diklasifikasikan dengan hasil yang didapatkan oleh metode DP setelah proses klasifikasi dilakukan. Untuk lebih jelasnya, ilustrasi kuisioner dapat dilihat pada Gambar 2.



Gambar 2. Ilustrasi Kuisioner Pengujian

Setelah dilakukan kuisioner, jumlah responden yang memilih hasil ekstraksi DP dengan hasil ekstraksi DP + klasifikasi akan dirata-rata kemudian dibandingkan. Hasil uji kuisioner untuk masing-masing produk dapat dilihat dalam Tabel 1.

Tabel 1.*Tabel Hasil Uji Kuisioner*

Produk	***DP Before Classification***	***DP After Classification***
Produk 1	0.3	0.7
Produk 2	0.2	0.8
Produk 3	0.4	0.6
Produk 4	0.3	0.7
Produk 5	0.4	0.6
Produk 6	0.4	0.6
Produk 7	0.3	0.7
Produk 8	0.3	0.7
Produk 9	0.4	0.6
Produk 10	0.2	0.8

Dari hasil uji coba kuisioner yang dilakukan pada scenario uji coba pertama, dapat disimpulkan bahwa pengkLasifikasian fitur produk sangat membantu *user* dalam menemukan fitur produk yang sebenarnya, Seperti dalam contoh pada ilustrasi Gambar 2 di atas, *user* akan lebih mudah memahami bahwa kalimat tersebut mengomentari *memory* dari produk *smartphone*, dibandingkan dengan jika fitur yang terekstrak adalah kata 8GB.

Sedangkan untuk uji *variable*, terdapat 2 *variabel* yang akan diuji yaitu *precision* dan *recall*. Kedua variable ini menyatakan seberapa besar akurasi metode *local point-wise mutual information* dalam mengklasifikasikan fitur produk dibandingkan dengan metode klasifikasi lainnya.

Uji coba dilakukan dengan memanfaatkan data acuan yang didapatkan dari kuisioner yang disebar kepada 30 *user* sebelumnya. Dalam kuisioner tersebut, *user* diminta memilih salah satu dari kelima kategori yang disediakan. *User* harus memilih salah satu kategori yang menurut dia paling merepresentasikan fitur produk yang sedang dikomentari oleh reviewer. Tabel 2 dan 3 menunjukkan hasil yang didapat dari hasil pengujian yang dilakukan.

Tabel 2.*Tabel Hasil Uji Precision*

Design	**Batery**	**Memory**	**Camera**	**Prize**	**Rata-rata**
0.78	0.81	0.83	0.69	0.97	0.82

Tabel 3.*Tabel Hasil Uji Recall*

Design	**Batery**	**Memory**	**Camera**	**Prize**	**Rata-rata**
0.88	0.85	0.85	0.83	0.98	0.88

Dari hasil uji variable *precision* dan *recall* dapat disimpulkan bahwa metode yang diusulkan mampu mengklasifikasikan fitur ke dalam kategori-kategoriyang disediakan dengan cukup baik. Terbukti dengan didapatkannya rata-rata *precision* dan *recall* di atas 80%.

5.KESIMPULAN

Proses klasifikasi fitur produk ternyata memang diperlukan untuk memudahkan *user* dalam menganalisa kelebihan dan kekurangan dari suatu produk. Hal itu dibuktikan dari hasil kuisioner yang menunjukkan bahwa sebagian besar *user* lebih menyukai hasil ektraksi setelah diklasifikasi dibandingkan dengan yang sebelumnya.

Dari penelitian ini juga dapat disimpulkan bahwa metode *Local Pointwise Mutual Information* yang diusulkan mampu mengatasi masalah ketergantungan *domain* yang seringkali terjadi dalam proses pengklasifikasian target opini. Hal ini dibuktian dengan nilai *precision* dan *recall* yang menunjukkan angka di atas 0.8.

6. DAFTAR PUSTAKA

[1] M. Hu and B. Liu. “Mining and Summarizing Customer Reviews”. Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD-2004), 8, pp. 168–174, 2004.

[2] Qiu, Guang., Bing, Liu., Jiajun Bu and Chun Chen. “Expanding *Domain* Sentiment Lexicon through Double Propagation”. In Proceedings of IJCAI, 2009.

[3] Azhar, Yufis, Agus Zainal Arifin, and Diana Purwitasari. "Otomatisasi Perbandingan Produk Berdasarkan Bobot Fitur pada Teks Opini." Jurnal Ilmu Komputer 6.2, 2013.

[4] Rozi, Fahrur, et al. "Pelabelan Klaster Fitur Secara Otomatis pada Perbandingan Review Produk." Jurnal Teknologi Informasi dan Ilmu Komputer 1.2, 2015.

[5] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.